

On randomness reduction in the Johnson-Lindenstrauss lemma

Paweł Wolff ^{*†}

Abstract

A refinement of so-called *fast Johnson-Lindenstrauss transform* (Ailon and Chazelle [2], Matoušek [17]) is proposed. While it preserves the time efficiency and simplicity of implementation of the original construction, it reduces randomness used to generate the random transformation. In the analysis of the construction two auxiliary results are established which might be of independent interest: a Bernstein-type inequality for a sum of a random sample from a family of independent random variables and a normal approximation result for such a sum.

Keywords: Johnson-Lindenstrauss lemma; Bernstein inequality; sampling without replacement; normal approximation

1 Introduction

The Johnson-Lindenstrauss lemma [15] is the following fact, which might appear quite surprising at the first sight:

Theorem 1. *Let $\varepsilon \in (0, 1)$, \mathcal{X} be an N -point subset of ℓ_2^n and $d \geq C \frac{\log N}{\varepsilon^2}$, where $C > 0$ is some universal constant. Then there exists a (linear) mapping $f: \ell_2^n \rightarrow \ell_2^d$ such that*

$$\forall_{x, y \in \mathcal{X}} \quad (1 - \varepsilon) \|x - y\|_2 \leq \|f(x) - f(y)\|_2 \leq (1 + \varepsilon) \|x - y\|_2. \quad (1)$$

Despite the original motivation for the Johnson-Lindenstrauss, it quickly became clear that this fact is of great importance in applications, especially in designing of algorithms which process high dimensional data (see e.g. [13, 1] and references therein). For this reason, lots of application-oriented variants of the above result appeared quite recently, e.g. [1, 2, 17, 3, 10, 21, 16]. In this paper we propose a refinement of the results of Ailon and Chazelle [2] and Matoušek [17]. In order to put our work into the context, we briefly sketch basic ideas behind, say, now classical proofs of the Johnson-Lindenstrauss lemma and comment on the papers [2] and [17] in some more details.

^{*}Institute of Mathematics. University of Warsaw. Banacha 2. 02-097 Warszawa. POLAND. Email: pwolff@mimuw.edu.pl

[†]Research partially supported by MNiSW Grant no. N N201 397437

Most of the known proofs of the Johnson-Lindenstrauss lemma provides the existence of the map f by drawing it according to some probability distribution on the space of $d \times n$ matrices and showing that it satisfies (1) with positive probability. The original proof of the Johnson-Lindenstrauss lemma [15] takes the map f to be an orthogonal projection onto a random d -dimensional subspace of ℓ_2^n (a random subspace means here a subspace drawn according to the normalized Haar measure on the Grassmannian $G_{d,n}$). It turns out (by means of a concentration inequality or, as originally, isoperimetry) that whenever $d \geq \frac{K}{\varepsilon^2}$ for a given constant $K > 0$, probability that f maps any fixed vector $v \in \ell_2^n$ onto a vector of the length $(1 \pm \varepsilon)\sqrt{d/n} \|v\|_2$ is at least $1 - C \exp(-cK)$, where $c, C > 0$ are universal constants. Taking K of order $\log N$ ensures that the failure probability is less than N^{-2} thus taking the union bound over $\binom{N}{2}$ vectors $v = x - y$ ($x, y \in \mathcal{X}$) the probability that the map $\sqrt{n/d}f$ fails (1) is less than $1/2$. Instead using random orthogonal projections one can use a random matrix which entries are i.i.d. Gaussian random variables [9] or the properly normalized matrix of independent random signs [1]. In each of these cases, however, the time complexity of evaluating $f(x)$ for a single point $x \in \mathcal{X}$ is $O(nd) = O(n \log N/\varepsilon^2)$ which sometimes is too much for practical applications. Also, the amount of randomness (measured in number of random bits, i.e. unbiased coin tosses) required to generate f is $O(nd)$.

Ailon and Chazelle [2] proposed a construction of a random map f which they called a *fast Johnson-Lindenstrauss transform* for the reason that in a wide range of parameters (basically N vs. n) it is computationally more efficient than the constructions described above. Assuming n is a power of 2, they take $f = PHD$ where D is an $n \times n$ diagonal matrix of random signs, H is the matrix of the Walsh-Hadamard transform on ℓ_2^n normalized by the factor $1/\sqrt{n}$ (so that H is an orthogonal matrix with all entries being $\pm 1/\sqrt{n}$), and P is some sparse random $d \times n$ matrix. The transformation HD is an isometry on ℓ_2^n and it can be shown that with probability close to 1 it maps any fixed unit vector $u \in \ell_2^n$ onto a vector $v \in \ell_2^n$ with small ℓ_∞ -norm. More precisely, for any constant $C > 0$, if $\|u\|_2 = 1$ and $V = HDu$, then with probability at least $1 - N^{-C}$,

$$\|V\|_\infty \leq C' \frac{\sqrt{\log N}}{\sqrt{n}}, \quad (2)$$

where $C' = C'(C) > 0$ is a constant depending on C only. This property is essential for the construction of the matrix P which is as follows: fix $q \in (0, 1]$ and set P to be a matrix of independent random entries, each entry equals 0 with probability $1 - q$ and with probability q is $\mathcal{N}(0, 1/(dq))$ random variable. It turns out that for any fixed unit vector $v \in \ell_2^n$ satisfying $\|v\|_\infty \leq c\sqrt{q}/\sqrt{\log(N/\varepsilon)}$, the probability that $1 - \varepsilon \leq \|Pv\|_2 \leq 1 + \varepsilon$ is at least $1 - N^{-C(c)}$. Together with (2) it implies that the map f will work whenever $q \geq C(\log N) \log(N/\varepsilon)/n$. This means the expected time of applying P to a single vector is $O(dqn) = O(\log^3 N/\varepsilon^2)$ (here we assume $\log(1/\varepsilon) = O(\log N)$). Since the transformation HD can be applied in time $O(n \log n)$ using the Fast Fourier Transform over the group $(\mathbb{Z}_2)^n$, this construction beats the previous approaches in terms of time complexity whenever $\log N = o(n^{1/2})$ and $\log N = \omega(\log n)$.

Since the usage of Gaussian random variables in the matrix P generally causes some extra technical problems in a practical implementation, Matoušek [17] refined the result of

Ailon and Chazelle replacing Gaussian r.v.'s with random signs (Bernoulli ± 1 r.v.'s). Also, in both papers, a similar property for the map f as a map from ℓ_2^n into ℓ_1^d was proved.

Generating the matrix P described above requires roughly $nd \log_2(1/q)$ random bits. This can be significantly reduced if we slightly change the distribution from which P is sampled. Instead of fully independent entries, let P now have only independent rows, and within each row we choose $k = nq$ entries at random (without replacement) in which we put a random sign. The remaining entries are zeros. This can be done using $O(k \log n) = O(\log^2 N \log n)$ random bits per row (see Section 3 for details). Additionally, in the matrix D it is enough to have only $O(\log N)$ -independent Bernoulli ± 1 random variables which can be modeled on the probability space $\{0, 1\}^{O(\log N \log n)}$ (see e.g. [4, Proposition 6.5] or [18, Chapter 7.6, Theorem 8]). Therefore, in total we use $O(\log^3 N \log n / \varepsilon^2)$ random bits and keep the computational efficiency and easiness of practical implementation of the constructions from [2] and [17].

The probabilistic analysis, similarly to the one done by Matoušek in [17], relies on tail estimates for sums of random variables. Here, however, these random variables are not fully independent. The main tools we established to perform the analysis is a Bernstein-type inequality and the L^1 Berry-Esseen bound for a sum of a random sample from a family of independent random variables. Although these results are not entirely new (see the comments following Theorem 7 and Theorem 13), we believe they still might be of some interest. Also, having potential applications of the result in mind, we provide explicit and reasonable numerical constants in estimates of parameters of our construction.

To finish the introduction, let us note that recently couple of other results in this area appeared, see [16] and references therein. Although these results beat our in terms of amount of randomness, the methods used there are (at least in part) quite different from ours and do not seem to work in the case of embedding into ℓ_1 .

2 Notation

Throughout the rest of the paper we use the following notation. Let $\varepsilon \in (0, 1)$, $\delta \in (0, \frac{1}{2})$ and a positive integer n be fixed parameters. Our goal is to construct a random linear map f which acts from ℓ_2^n to a space (ℓ_2 or ℓ_1) of a smaller dimension and satisfies the following property: for any fixed $u \in \ell_2^n$,

$$\mathbb{P}((1 - \varepsilon) \|u\|_2 \leq \|f(u)\| \leq (1 + \varepsilon) \|u\|_2) \geq 1 - 2\delta.$$

Assume n is a power of 2 (if necessary we augment a vector u with zeros). Let d and $k \leq n$ be positive integers to be specified later. We shall consider the following families of random variables:

- $\beta_1, \beta_2, \dots, \beta_n$ are symmetric ± 1 random variables and l -independent with $l := 2 \lceil \log(n/\delta) \rceil$, i.e. any l of these random variables are independent. If $l > n$ then β_1, \dots, β_n are just independent.
- $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ are independent symmetric ± 1 Bernoulli random variables. For $i = 1, \dots, d$, $(\varepsilon_{i,1}, \varepsilon_{i,2}, \dots, \varepsilon_{i,n})$ are independent copies of $(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)$.

- $\xi_1, \xi_2, \dots, \xi_n$ are 0–1 random variables such that the distribution of a random sets $\{j: \xi_j = 1\} \subset \{1, \dots, n\}$ is uniform over all subsets of $\{1, \dots, n\}$ with cardinality k . In the other words, for any $J \subset \{1, \dots, n\}$ with cardinality k , $\mathbb{P}(\{j: \xi_j = 1\} = J) = 1/\binom{n}{k}$. For $i = 1, \dots, d$, $(\xi_{i,1}, \xi_{i,2}, \dots, \xi_{i,n})$ are independent copies of $(\xi_1, \xi_2, \dots, \xi_n)$.

Moreover, all the families of random variables are independent, that is $\sigma(\beta_1, \dots, \beta_n)$, $\sigma(\varepsilon_1, \dots, \varepsilon_n)$, $\sigma(\varepsilon_{1,1}, \dots, \varepsilon_{1,n}), \dots, \sigma(\varepsilon_{d,1}, \dots, \varepsilon_{d,n})$, $\sigma(\xi_1, \dots, \xi_n)$, $\sigma(\xi_{1,1}, \dots, \xi_{1,n}), \dots, \sigma(\xi_{d,1}, \dots, \xi_{d,n})$ are independent.

For $q \in \{1, 2\}$ we define a random linear map $f_q: \ell_2^n \rightarrow \ell_q^d$ as follows:

$$f_q = \frac{1}{d^{1/q}} P H D \quad (3)$$

where

$$D = \begin{pmatrix} \beta_1 & & 0 \\ & \ddots & \\ 0 & & \beta_n \end{pmatrix},$$

$$P = \sqrt{\frac{n}{k}} \begin{pmatrix} \xi_{1,1}\varepsilon_{1,1} & \cdots & \xi_{1,n}\varepsilon_{1,n} \\ \vdots & \ddots & \vdots \\ \xi_{d,1}\varepsilon_{d,1} & \cdots & \xi_{d,n}\varepsilon_{d,n} \end{pmatrix}$$

and $H = H_n$ is the normalized Walsh-Hadamard matrix of size $n \times n$, that is the orthogonal matrix defined by the following recursive formula:

$$H_n = \frac{1}{\sqrt{2}} \begin{pmatrix} H_{n/2} & H_{n/2} \\ H_{n/2} & -H_{n/2} \end{pmatrix} \quad \text{if } n > 1,$$

$$H_1 = (1).$$

The function \log stands for the natural logarithm. We write $\|\cdot\|_q$ for the ℓ_q norm ($1 \leq q \leq \infty$).

3 The results

The main result of this paper states that the random transformation $f_q: \ell_2^n \rightarrow \ell_q^d$ ($q = 1, 2$) defined above with probability close to 1 almost preserves the norm of any fixed vector, provided certain bounds on the parameters d and k hold.

Theorem 2 (the ℓ_2 case). *Assume n is a power of 2,*

$$d \geq 1.55 \frac{(1 + 2\varepsilon)^2}{\varepsilon^2} \log(3/\delta) \quad \text{and} \quad k \geq \max\left(\frac{8e}{3} \log(6d/\delta), 20e\right) \log(2n/\delta).$$

Then the random linear transformation f_2 as defined in (3) satisfies

$$\forall u \in \ell_2^n \quad \mathbb{P}((1 + \varepsilon)^{-1} \|u\|_2 \leq \|f_2(u)\|_2 \leq (1 + \varepsilon) \|u\|_2) \geq 1 - 2\delta$$

provided $k \leq n$.

Theorem 3. Assume n is a power of 2. For any constant $\kappa \in (0, 1)$, assume

$$d \geq \frac{\pi + \sqrt{\pi/2} \frac{8}{3} \kappa \varepsilon}{\kappa^2 \varepsilon^2} \log(2/\delta) \quad \text{and} \quad k \geq \max \left(\frac{9\pi e}{4(1-\kappa)^2 \varepsilon^2}, 20e \right) \log(2n/\delta).$$

Then the random linear transformation f_1 as defined in (3) satisfies

$$\forall u \in \ell_2^n \quad \mathbb{P}((1 - \varepsilon) \|u\|_2 \leq \sqrt{\pi/2} \|f_1(u)\|_1 \leq (1 + \varepsilon) \|u\|_2) \geq 1 - 2\delta$$

provided $k \leq n$.

The typical situation in which these results are applied is the one mentioned in the introduction: we have N points in ℓ_2^n (n is a power of 2) and we want to embed them into a space of (possibly much lower) dimension d with distortion $1 + \varepsilon$. To this end we choose an embedding at random, as specified in Theorem 2 or 3. Assuming we want the embedding to work with probability at least $1 - p$, where $p \in (0, 1)$ is a fixed parameter, we take $\delta = p/N^2$ and apply the union bound over $\binom{N}{2}$ vectors being differences of pairs of points to get that the embedding fails to have distortion $1 + \varepsilon$ with probability at most $\binom{N}{2} 2\delta < p$.

In the case of embedding into ℓ_2 ,

$$d = \left\lceil 1.55 \frac{(1 + 2\varepsilon)^2}{\varepsilon^2} \log \left(\frac{3N^2}{p} \right) \right\rceil, \quad k = \left\lceil \max \left(7.25 \log \left(\frac{6dN^2}{p} \right), 55 \right) \log \left(\frac{2nN^2}{p} \right) \right\rceil$$

satisfy the hypothesis of Theorem 2 unless $k > n$. If indeed $k > n$, or even $k > n/3$, then one can use the construction of Achlioptas [1] which provides a random embedding into ℓ_2^d with d similar to ours, roughly with constant 1 instead of 1.55. The embedding is given by a $d \times n$ matrix with entries being independent random variables assuming values $1, 0, -1$ with respective probabilities $\frac{1}{6}, \frac{2}{3}, \frac{1}{6}$. See [1] for details.

Therefore, in what follows, we assume $k \leq n/3$. The construction of the random embedding f_2 is actually the matter of constructing the random variables β_1, \dots, β_n and $\xi_{1,1\varepsilon_{1,1}}, \dots, \xi_{1,n\varepsilon_{1,n}}, \dots, \xi_{d,1\varepsilon_{d,1}}, \dots, \xi_{d,n\varepsilon_{d,n}}$ on a sample space $\{0, 1\}^r$ with the uniform probability, where r will be the number of random bits used. Due to the construction of Alon, Babai and Itai [4, Proposition 6.5], the l -independent (now $l = 2\lceil \log(N^2 n/p) \rceil$) symmetric ± 1 random variables β_1, \dots, β_n can be constructed on the uniform sample space $\{0, 1\}^{(\log_2 n + 1)l/2 + 1}$, thus $O((\log n) \log(Nn))$ random bits suffice. (Moreover, given an element of the sample space, their construction allows to compute the sequence β_1, \dots, β_n in time $O(\ln \log n) = O(n(\log(Nn)) \log n)$.) For a given $i \in \{1, \dots, d\}$, $\xi_{i,1\varepsilon_{i,1}}, \dots, \xi_{i,n\varepsilon_{i,n}}$ can be constructed as follows:

```

J ← ∅
while #J < k do
  using log2 n random bits sample an index j uniformly in {1, ..., n}
  if j ∉ J then
    J ← J ∪ {j}
  end if
end while

```

Thus J becomes a random subset of $\{1, \dots, n\}$ of size k . Now set $\xi_{i,j} = 1$ for $j \in J$ and $\xi_{i,j} = 0$ for $j \notin J$ and using k random bits sample $\varepsilon_{i,j}$ for each $j \in J$. The total number of random bits used is $k + T_i \log_2 n$, where T_i is the number of iterations made by the **while** loop. Note that T_i is a sum of k independent geometric random variables with subsequent success probabilities $1, \frac{n-1}{n}, \dots, \frac{n-k+1}{n}$ (the success is sampling j not yet contained in J). Since $k \leq n/3$, a rough estimate gives $\mathbb{E}T_i \leq \frac{3}{2}k$ and $\text{Var}(T_i) \leq \frac{3}{4}k$. Proceeding in the same way for each row, we construct the matrix P by using $dk + T \log_2 n$ random bits, where $T = T_1 + \dots + T_d$ and T_i are independent. Thus $\mathbb{E}T \leq \frac{3}{2}kd$ and $\text{Var}(T) \leq \frac{3}{4}kd$. By the Chebyshev inequality, for $\lambda > 0$,

$$\mathbb{P}\left(T \geq \frac{3}{2}kd + \lambda\sqrt{\frac{3}{4}kd}\right) \leq \mathbb{P}\left(T \geq \mathbb{E}T + \lambda\sqrt{\text{Var}(T)}\right) \leq \lambda^{-2}$$

hence we see that with probability close to 1, T does not exceed some constant times kd . (Actually one can derive much stronger exponential tail estimate for T , but it is not essential here.) Overall, the whole construction uses $O((\log n) \log(Nn) + dk \log n)$ random bits to generate the random embedding f_2 . Assuming p is fixed and $\log n = O(\log N)$ and $\log(1/\varepsilon) = O(\log N)$, we have $d = O(\varepsilon^{-2} \log N)$, $k = O((\log N)^2)$ and the number of random bits used is $O(\varepsilon^{-2}(\log N)^3 \log n)$. The time complexity of applying f_2 to a single point is $O(dk + n \log n) = O(\varepsilon^{-2}(\log N)^3 + n \log n)$, using the Fast Fourier Transform.

The case of embedding into ℓ_1 is similar, since the random transformation f_1 used has the same structure as f_2 . Let us note, however, that this time the requirement $k \leq n$ in Theorem 3 is in some sense restrictive and cannot be circumvented as previously. For any $\kappa \in (0, 1)$,

$$d = \left\lceil \frac{3.15 + 3.4\kappa\varepsilon}{\kappa^2\varepsilon^2} \log\left(\frac{2N^2}{p}\right) \right\rceil, \quad k = \left\lceil \max\left(\frac{19.3}{(1-\kappa^2)\varepsilon^2}, 55\right) \log\left(\frac{2nN^2}{p}\right) \right\rceil$$

satisfy the hypothesis of Theorem 3 as long as $k \leq n$. If $k \leq n/3$ (or some other fraction of n), we may proceed with the same algorithm of sampling the matrix P . In such case, the total number of random bits used to generate f_1 is $O(\log n(\log(Nn)) + dk \log n)$ which is $O(\varepsilon^{-4}(\log N)^2 \log n)$, again assuming $\log n = O(\log N)$, $\log(1/\varepsilon) = O(\log N)$. If k is between $n/3$ and n , we may still get an embedding into ℓ_1 . Since the sparsity of the matrix P is rather poor, one can set k to be n while possibly increasing the parameter $\kappa \in (0, 1)$ in order to slightly reduce the target dimension d . Of course, having $k = n$ there is no need to sample $\xi_{i,j}$ as all are equal 1.

Finally, if $k > n$ for all $\kappa \in (0, 1)$, which happens when $\log N = \Omega(\varepsilon^2 n)$, then even for $k = n$ Theorem 3 cannot guarantee the random map f_1 is an embedding of a set of N points in ℓ_2^n into ℓ_1^d with distortion $1 + \varepsilon$. The reason for that is the distortion of f_1 depends on the error of approximation

$$\mathbb{E}\left|\sum_{j=1}^n v_j \varepsilon_j\right| \approx \mathbb{E}|G| = \sqrt{2/\pi},$$

where $\sum_{j=1}^n v_j^2 = 1$ and G is a standard Gaussian random variable (see the next section for details). If one knows that all $|v_j|$ are small, then the distribution of $\sum_{j=1}^n v_j \varepsilon_j$ is close

to the standard Gaussian. (A suitable version of that statement, working also in the case $k < n$ and thus dealing with the sums $\sum_{j=1}^n v_j \xi_j \varepsilon_j$, is formulated as Theorem 13.) On the other hand, if, say, one of v_j 's is close to ± 1 , then the error of the above approximation is of constant order rather than of order of ε . Since v_1, \dots, v_n are in fact the coordinates of a vector HDu , where u is one out of N arbitrary unit vectors in ℓ_2^n , the bound on $\max_j |v_j|$ that can be guaranteed (with non-negligible probability and randomness involved is due to the matrix D) is of order $\min \left\{ \sqrt{\frac{\log(Nn)}{\sqrt{n}}}, 1 \right\}$ (see Proposition 5 for the precise statement). This bound becomes trivial if $\log N = \Omega(n)$, but one cannot expect anything essentially better. Indeed, take a set \mathcal{X} of $N = 2^n$ unit vectors $(\pm 1/\sqrt{n}, \dots, \pm 1/\sqrt{n})$ in ℓ_2^n . Then whatever instance of the matrix D we consider, there is always a vector $u \in \mathcal{X}$ such that $HDu = (1, 0, \dots, 0)$. In the other words, $\mathbb{P}(\exists u \in \mathcal{X} \|HDu\|_\infty = 1) = 1$.

4 Proofs

The proof of Theorems 2 and 3 consists of four steps, which we outline below:

1. We show that for any unit vector $u \in \ell_2^n$, the random vector $V = HDu$ has typically the ℓ_∞ norm less than $C\sqrt{\log(n/\delta)}/\sqrt{n}$.
2. If a unit vector $v \in \ell_2^n$ has small ℓ_∞ norm, then each coordinate of the random vector $W = (W_1, \dots, W_d) = Pv$, which is distributed as the sum $\sqrt{n/k} \sum_{j=1}^n \xi_j \varepsilon_j v_j$, is well concentrated. More precisely, in the case of embedding into ℓ_2 space we shall show that W_i^2 is well concentrated around its mean $\mathbb{E}W_i^2$. In the case of ℓ_1 embedding, we show the concentration of $|W_i|$ around $\mathbb{E}|W_i|$. In both cases we use a version of Bernstein inequality for a sum of a random sample from a family of independent random variables (Theorem 7).
3. In the case of ℓ_2 embedding we note that $\mathbb{E}W_i^2$ depends only on the length of the vector v , and if $\|v\|_2 = 1$, then $\mathbb{E}W_i^2 = 1$. Since it is no longer true for $\mathbb{E}|W_i|$, in the ℓ_1 case we shall use a normal approximation of the distribution of W_i (Theorem 13) in order to show that $\mathbb{E}|W_i|$ is close to $\sqrt{2/\pi}$.
4. If the random vector W has all its coordinates well concentrated around a certain value then $\frac{1}{d} \|W\|_2^2$ or $\frac{1}{d} \|W\|_1$ is well concentrated.

In the subsequent sections we elaborate on each of these steps in detail.

4.1 Random signs and the Walsh-Hadamard transform

Assume $u \in \ell_2^n$ is a unit vector and let $V = (V_1, \dots, V_n) = HDu$. Since $H: \ell_2^n \rightarrow \ell_2^n$ is an isometry, $\|V\|_2 = 1$ a.s. Also, H has all entries $\pm 1/\sqrt{n}$ thus each coordinate V_i has the distribution of a random variable $\frac{1}{\sqrt{n}} \sum_{j=1}^n \beta_j x_j$, where $x_j = \pm u_j$ (the particular choice of the signs depends on i), in particular, $\sum_{j=1}^n x_j^2 = 1$.

Lemma 4. *If $\sum_{j=1}^n x_j^2 = 1$ and $S = \sum_{j=1}^n \beta_j x_j$, then*

$$\mathbb{P}\left(|S| \geq \sqrt{2e \log(2n/\delta)}\right) \leq \delta/n.$$

Proof. Recall that β_1, \dots, β_n are l -independent random variables with $l = 2\lceil \log(n/\delta) \rceil$. Therefore, for (fully) independent Bernoulli sequence $\varepsilon_j = \pm 1$, we have

$$\mathbb{E}S^l = \mathbb{E}\left(\sum_{j=1}^n \varepsilon_j x_j\right)^l$$

(just expand both sides, use linearity of expectation and note that each term involves the expectation of the product of at most $\min(l, n)$ distinct β_j 's or ε_j 's). The classical Khintchine inequality states

$$\left(\mathbb{E}\left|\sum \varepsilon_j x_j\right|^p\right)^{1/p} \leq C_p \left(\sum x_j^2\right)^{1/2}$$

for any $p \geq 2$ with some constant C_p depending on p only. It follows e.g. from classical hypercontractive estimates for Bernoulli random variables (see [6]) that the inequality holds with $C_p = \sqrt{p-1}$. Taking $p := l = 2\lceil \log(1/\delta) + \log n \rceil$, we thus have

$$(\mathbb{E}S^p)^{1/p} \leq \sqrt{p-1} < \sqrt{2 \log(n/\delta) + 1} < \sqrt{2 \log(2n/\delta)}$$

which combined with the Markov inequality

$$\mathbb{P}\left(|S| \geq \sqrt{e}(\mathbb{E}|S|^p)^{1/p}\right) \leq e^{-p/2} \leq \delta/n$$

finishes the proof. \square

Taking the union bound over all coordinates of V we immediately arrive with the following

Proposition 5. *Let $u \in \ell_2^n$ be a unit vector and let $V = HDu$. Then*

$$\mathbb{P}\left(\|V\|_\infty \geq \frac{\sqrt{2e \log(2n/\delta)}}{\sqrt{n}}\right) \leq \delta.$$

4.2 Bernstein inequality for a random sample from independent r.v.'s

First, let us recall the classical Bernstein inequality. Let Y_1, Y_2, \dots, Y_n be independent random variables with $\mathbb{E}Y_i = 0$. We assume all moments of Y_i 's are finite and for some constant $M > 0$,

$$\mathbb{E}|Y_i|^p \leq \frac{p!}{2} \sigma_i^2 M^{p-2}, \quad \text{for any integer } p \geq 2. \quad (4)$$

The classical inequality of Bernstein provides the estimate for the tail of the sum $Y_1 + \dots + Y_n$:

Theorem 6. Let $S = Y_1 + \dots + Y_n$ with Y_i 's satisfying (4) and set $\sigma^2 = \sum_{i=1}^n \sigma_i^2$. Then for all $s > 0$,

$$\mathbb{P}(S \geq s) \leq \exp\left(-\frac{s^2}{2\sigma^2 + 2Ms}\right) \quad \text{and} \quad \mathbb{P}(S \leq -s) \leq \exp\left(-\frac{s^2}{2\sigma^2 + 2Ms}\right).$$

Beside the classical Bernstein inequality, we shall also use its variant for a sum of a random sample of k out of n random variables Y_1, \dots, Y_n , that is for the sum $\sum_{i=1}^n \xi_i Y_i$.

Theorem 7. Let $S = \sum_{i=1}^n \xi_i Y_i$ with Y_i 's satisfying (4) and set $\sigma^2 = \sum_{i=1}^n \sigma_i^2$. Then for all $s > 0$,

$$\mathbb{P}(S \geq s) \leq \exp\left(-\frac{s^2}{2\frac{k}{n}\sigma^2 + 2Ms}\right) \quad \text{and} \quad \mathbb{P}(S \leq -s) \leq \exp\left(-\frac{s^2}{2\frac{k}{n}\sigma^2 + 2Ms}\right).$$

(Recall, $\mathbb{P}(\xi_i = 1) = k/n$.)

We will need a simple

Lemma 8. For any $A \subseteq \{1, \dots, n\}$,

$$\mathbb{E} \prod_{i \in A} \xi_i \leq \prod_{i \in A} \mathbb{E} \xi_i = \left(\frac{k}{n}\right)^{\#A}.$$

Proof. If $\#A > k$ then $\mathbb{E} \prod_{i \in A} \xi_i = 0$, otherwise

$$\begin{aligned} \mathbb{E} \prod_{i \in A} \xi_i &= \mathbb{P}(\xi_i = 1 \text{ for each } i \in A) = \frac{\binom{n-\#A}{k-\#A}}{\binom{n}{k}} \\ &= \frac{k(k-1)\dots(k-\#A+1)}{n(n-1)\dots(n-\#A+1)} \leq \left(\frac{k}{n}\right)^{\#A}. \end{aligned}$$

□

Proof of Theorem 7. Except for using Lemma 8, the proof follows a standard proof of Bernstein inequality. We present the proof below for the sake of completeness.

First, for any i and $|t| < 1/M$,

$$\mathbb{E} e^{tY_i} = \sum_{k=0}^{\infty} \frac{t^k}{k!} \mathbb{E} Y_i^k \leq 1 + \frac{\sigma_i^2}{2} \sum_{k=2}^{\infty} |t|^k M^{k-2} \leq 1 + \frac{\sigma_i^2 t^2}{2(1-M|t|)}. \quad (5)$$

To estimate $\mathbb{E} e^{tS}$, we condition on $\mathcal{F} = \sigma(\xi_i : i = 1, \dots, n)$, use (5) and Lemma 8:

$$\begin{aligned} \mathbb{E} e^{tS} &= \mathbb{E} \prod_{i=1}^n \mathbb{E}[e^{t\xi_i Y_i} | \mathcal{F}] = \mathbb{E} \prod_{i=1}^n (1 - \xi_i + \xi_i \mathbb{E} e^{tY_i}) \\ &\leq \mathbb{E} \prod_{i=1}^n \left(1 + \xi_i \frac{\sigma_i^2 t^2}{2(1-M|t|)}\right) \leq \prod_{i=1}^n \left(1 + (\mathbb{E} \xi_i) \frac{\sigma_i^2 t^2}{2(1-M|t|)}\right) \\ &\leq \exp\left(\frac{k}{n} \frac{\sigma^2 t^2}{2(1-M|t|)}\right). \end{aligned}$$

One obtains the inequality for the tail probability $\mathbb{P}(S \geq s)$ by taking $t = \frac{s}{\frac{k}{n}\sigma^2 + Ms}$ and using Chebyshev's inequality. For the lower tail use $\mathbb{P}(S \leq -s) = \mathbb{P}(-S \geq s)$. \square

Remark 9. Since the random variables ξ_1, \dots, ξ_n are negatively associated (see [14]), the above result can be deduced (up to numerical constants) from a quite general comparison result of Shao [19]. See also the paper of Hoeffding [12] for related results.

We use Theorem 7 to obtain concentration for coordinates of the vector $W = Pv$, i.e.

$$W_i = \sqrt{\frac{n}{k}} \sum_{j=1}^n v_j \xi_{i,j} \varepsilon_{i,j} \quad \text{for } i = 1, \dots, d.$$

Proposition 10. Assume $v \in \ell_2^n$, $\|v\|_2 = 1$, $\|v\|_\infty \leq \alpha$ and let $W = Pv$. Then for $i = 1, \dots, d$ and any $s > 0$,

$$\mathbb{P}(|W_i| \geq s) \leq 2 \exp \left(-\frac{s^2}{2 + \frac{2}{3}(n/k)^{1/2} \alpha s} \right).$$

Proof. Fix any $i \in \{1, \dots, d\}$ and set $Y_j = (n/k)^{1/2} \varepsilon_{i,j} v_j$. With $\sigma_j^2 = (n/k) v_j^2$ and $M = (n/k)^{1/2} \alpha/3$, the condition (4) is satisfied. Since $W_i = \sum_{j=1}^n \xi_{i,j} Y_j$ and $\sigma^2 = \sum_{j=1}^n \sigma_j^2 = n/k$, Theorem 7 provides the desired bound on $\mathbb{P}(|W_i| \geq s)$. \square

For the sake of providing good numerical constants, beside the tail estimates established above we estimate a few first even moments of W_i under the additional assumption

$$\frac{n}{k} \alpha^2 \leq r_0 := \frac{1}{10}. \quad (6)$$

Lemma 11. Under the assumptions of Proposition 10, if (6) holds then

$$\mathbb{E}W_i^4 \leq 3.1, \quad \mathbb{E}W_i^6 \leq 17, \quad \mathbb{E}W_i^8 \leq 127, \quad \mathbb{E}W_i^{10} \leq 1283.$$

Proof. For $q = 2, 3, 4, 5$, write

$$\mathbb{E}W_i^{2q} = \left(\frac{n}{k}\right)^q \mathbb{E} \left(\sum_{j=1}^n v_j \xi_j \varepsilon_j \right)^{2q}$$

and expand the right hand side. By the symmetry and independence of ε_j 's, all the terms in the expansion involving odd powers vanish. Using the fact that for any integer $q_1 \geq 1$,

$$\sum_{j=1}^n v_j^{2q_1} \leq \left(\sum_{j=1}^n v_j^2 \right) \|v\|_\infty^{2(q_1-1)} \leq \alpha^{2(q_1-1)},$$

and Lemma 8 we thus obtain

$$\begin{aligned}\mathbb{E}W_i^4 &= \left(\frac{n}{k}\right)^2 \mathbb{E} \left(\sum_{j=1}^n v_j \xi_j \varepsilon_j \right)^4 = \left(\frac{n}{k}\right)^2 \left(\sum_{j=1}^n v_j^4 \mathbb{E} \xi_j + 3 \sum_{j_1 \neq j_2} v_{j_1}^2 v_{j_2}^2 \mathbb{E} \xi_{j_1} \xi_{j_2} \right) \\ &\leq \left(\frac{n}{k}\right)^2 \alpha^2 (k/n) + 3 \left(\frac{n}{k}\right)^2 \left(\sum_{j_1, j_2} v_{j_1}^2 v_{j_2}^2 \right) (k/n)^2 = \frac{n}{k} \alpha^2 + 3 \leq r_0 + 3.\end{aligned}$$

Similarly,

$$\begin{aligned}\mathbb{E}W_i^6 &\leq \left(\frac{n}{k}\right)^3 \left((k/n) \sum_j v_j^6 + \binom{6}{4\ 2} (k/n)^2 \sum_{j_1 \neq j_2} v_{j_1}^4 v_{j_2}^2 + \frac{\binom{6}{2\ 2\ 2}}{3!} (k/n)^3 \sum_{\substack{j_1, j_2, j_3 \\ \text{(distinct)}}} v_{j_1}^2 v_{j_2}^2 v_{j_3}^2 \right) \\ &\leq (n/k)^2 \alpha^4 + 15(n/k) \alpha^2 + 15 \leq r_0^2 + 15r_0 + 15,\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}W_i^8 &\leq \left(\frac{n}{k}\right)^4 \left((k/n) \sum_j v_j^8 + \binom{8}{6\ 2} (k/n)^2 \sum_{j_1 \neq j_2} v_{j_1}^6 v_{j_2}^2 + \frac{\binom{8}{4\ 4}}{2!} (k/n)^2 \sum_{j_1 \neq j_2} v_{j_1}^4 v_{j_2}^4 \right. \\ &\quad \left. + \frac{\binom{8}{4\ 2\ 2}}{2!} (k/n)^3 \sum_{\substack{j_1, j_2, j_3 \\ \text{(distinct)}}} v_{j_1}^4 v_{j_2}^2 v_{j_3}^2 + \frac{\binom{8}{2\ 2\ 2\ 2}}{4!} (k/n)^4 \sum_{\substack{j_1, j_2, j_3, j_4 \\ \text{(distinct)}}} v_{j_1}^2 v_{j_2}^2 v_{j_3}^2 v_{j_4}^2 \right) \\ &\leq (n/k)^3 \alpha^6 + 28(n/k)^2 \alpha^4 + 35(n/k)^2 \alpha^4 + 210(n/k) \alpha^2 + 105 \\ &\leq r_0^3 + 28r_0^2 + 35r_0^2 + 210r_0 + 105,\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}W_i^{10} &\leq r_0^4 + \binom{10}{8\ 2} r_0^3 + \binom{10}{6\ 4} r_0^3 + \frac{\binom{10}{6\ 2\ 2}}{2!} r_0^2 + \frac{\binom{10}{4\ 4\ 2}}{2!} r_0^2 + \frac{\binom{10}{4\ 2\ 2\ 2}}{3!} r_0 + \frac{\binom{10}{2\ 2\ 2\ 2\ 2}}{5!} \\ &= r_0^4 + 45r_0^3 + 210r_0^3 + 630r_0^2 + 1575r_0^2 + 3150r_0 + 945.\end{aligned}$$

□

We shall use the following simple lemma to handle the deviation of W_i^2 or $|W_i|$ from their means.

Lemma 12. *Assume $Y \geq 0$ a.s., $a \geq 0$ and $\Phi: \mathbb{R}_+ \rightarrow \mathbb{R}_+$ is non-decreasing. Then*

$$\mathbb{E}\Phi(|Y - a|) \leq \mathbb{E}\Phi(Y) + \Phi(a).$$

Proof. Note, that $\Phi(|Y - a|)\mathbf{1}_{\{Y \geq a\}} \leq \Phi(Y)$ a.s. and $\Phi(|Y - a|)\mathbf{1}_{\{Y < a\}} \leq \Phi(a)$ a.s. Summing up both inequalities and taking the expectation concludes the proof. □

4.3 $\mathbb{E} \|Pv\|_2^2 = d$ and $\mathbb{E} \|Pv\|_1 \approx d\sqrt{2/\pi}$ by normal approximation

Let $v \in \ell_2^n$ be a unit vector and $W = Pv$. Note that W_1, \dots, W_d are independent and

$$\mathbb{E} W_i^2 = \frac{n}{k} \mathbb{E} \left(\sum_{j=1}^n v_j \xi_j \varepsilon_j \right)^2 = \frac{n}{k} \sum_j v_j^2 \mathbb{E} \xi_j = 1,$$

hence $\mathbb{E} \|W\|_2^2 = d$.

The case of ℓ_1 -norm is more delicate. In principle, $\mathbb{E}|W_i|$ depends on $v = (v_1, \dots, v_n)$. However under the assumptions of small ℓ_∞ -norm of v and k large, the distribution of W_i is approximately Gaussian and thus $\mathbb{E}|W_i|$ can be approximated by $\sqrt{2/\pi}$. To this end we establish a slightly more general result which can be regarded as L^1 Berry-Esseen bound for a random sample from a family of independent random variables.

Theorem 13. *Let $n \geq 2$, Y_i , $i = 1, 2, \dots, n$ be independent random variables and independent of (ξ_1, \dots, ξ_n) , satisfying $\mathbb{E} Y_i = 0$, $\sum_{i=1}^n \mathbb{E} Y_i^2 = n/k$ and having finite third moment. Denote $X_i = \xi_i Y_i$ and $S = \sum_{i=1}^n X_i$. Then the Wasserstein distance between the distribution of S and the standard normal distribution*

$$d_W(S, G) := \sup_{h \in \text{Lip}(1)} |\mathbb{E} h(S) - \mathbb{E} h(G)| \leq 3 \sum_{i=1}^n \mathbb{E} |X_i|^3,$$

where $G \sim \mathcal{N}(0, 1)$ and $\text{Lip}(1)$ is a set of 1-Lipschitz functions on \mathbb{R} . Moreover,

$$|\mathbb{E}|S| - \sqrt{2/\pi}| \leq \frac{3}{2} \sum_{i=1}^n \mathbb{E} |X_i|^3 = \frac{3}{2} \frac{k}{n} \sum_{i=1}^n \mathbb{E} |Y_i|^3. \quad (7)$$

In the literature there exist many related results, most of them concerning more general problem called combinatorial central limit theorem. However, the author was not able to find a result which implies (7) with a reasonable numerical constant. The combinatorial central limit theorem roughly states that $S_n = \sum_{i=1}^n Y_{i, \pi(i)}$ where $(X_{i,j})_{i,j \leq n}$ is a matrix of independent random variables having finite third moments and π is a random permutation of the set $\{1, 2, \dots, n\}$, independent from $X_{i,j}$'s, after proper normalization has the distribution close to standard normal. Taking the matrix $(X_{i,j})$, whose first k rows are independent copies of the random vector (Y_1, \dots, Y_n) and the remaining entries are zeros, boils down to the problem from Theorem 13.

For example, the result of Ho and Chen [11, Theorem 3.1] on the combinatorial CLT implies an estimate similar to (7) but asymptotically weaker. Bolthausen [5] proved an optimal error bound but only in the case of deterministic (X_{ij}) 's. Recently, Chen and Fang [8] proved the general version of combinatorial CLT with the optimal rate of normal approximation error. They bound the Kolmogorov distance, which is generally more difficult to handle in comparison to the Wasserstein distance. However, for our purposes the Wasserstein distance is better suited and moreover it is possible to obtain an estimate with a reasonable numerical constant.

As in the results on combinatorial CLT mentioned above, we employ Stein's method. Except for a few twists, we basically follow the reasoning presented in [7, Section 2] which

illustrates the usage of Stein's method in the most basic setting of sums of independent random variables.

Proof. It is enough to consider a 1-Lipschitz $h: \mathbb{R} \rightarrow \mathbb{R}$ which is piecewise continuously differentiable. As in [7, Section 2.1], consider the differential equation

$$f'(x) - xf(x) = h(x) - \mathbb{E}h(G) \quad (8)$$

whose solution is given by the formula

$$f(x) = e^{x^2/2} \int_{-\infty}^x (h(t) - \mathbb{E}h(G)) e^{-t^2/2} dt. \quad (9)$$

Note that f is C^1 and f' is piecewise continuously differentiable, so for any $a, b \in \mathbb{R}$, $|f'(a) - f'(b)| \leq |a - b| \|f''\|_\infty$. It turns out [7, Lemma 2.3] that $\|f\|_\infty \leq 2 \|h'\|_\infty \leq 2$, $\|f'\|_\infty \leq 4 \|h'\|_\infty \leq 4$ and $\|f''\|_\infty \leq 2 \|h'\|_\infty \leq 2$.

Putting S as x in (8) and taking the expectation we get

$$\mathbb{E}h(S) - \mathbb{E}h(G) = \mathbb{E}(f'(S) - Sf(S)).$$

Set $S^{(i)} = S - X_i$ and define

$$K_i(t) = \mathbb{E}X_i (\mathbf{1}_{\{0 \leq t \leq X_i\}} - \mathbf{1}_{\{X_i \leq t < 0\}}).$$

Note that $K_i(t) \geq 0$ for all $t \in \mathbb{R}$,

$$\int_{-\infty}^{\infty} K_i(t) dt = \mathbb{E}X_i^2 \quad \text{and} \quad \int_{-\infty}^{\infty} |t| K_i(t) dt = \frac{1}{2} \mathbb{E}|X_i|^3. \quad (10)$$

Since Y_i is mean-zero and independent of $\sigma(S^{(i)}, \xi_i)$, we have $\mathbb{E}X_i f(S^{(i)}) = \mathbb{E}Y_i \mathbb{E}\xi_i f(S^{(i)}) = 0$. Therefore

$$\begin{aligned} \mathbb{E}Sf(S) &= \sum_{i=1}^n \mathbb{E}X_i f(S) = \sum_{i=1}^n \mathbb{E}X_i (f(S) - f(S^{(i)})) \\ &= \sum_{i=1}^n \mathbb{E}X_i \int_0^{X_i} f'(S^{(i)} + t) dt = \sum_{i=1}^n \int_{-\infty}^{\infty} \mathbb{E}f'(S^{(i)} + t) X_i (\mathbf{1}_{\{0 \leq t \leq X_i\}} - \mathbf{1}_{\{X_i \leq t < 0\}}) dt \\ &= \sum_{i=1}^n \int_{-\infty}^{\infty} \mathbb{P}(\xi_i = 1) \mathbb{E}[f'(S^{(i)} + t) Y_i (\mathbf{1}_{\{0 \leq t \leq Y_i\}} - \mathbf{1}_{\{Y_i \leq t < 0\}}) | \xi_i = 1] dt \\ &= \sum_{i=1}^n \int_{-\infty}^{\infty} \frac{k}{n} \mathbb{E}[f'(S^{(i)} + t) | \xi_i = 1] \mathbb{E}Y_i (\mathbf{1}_{\{0 \leq t \leq Y_i\}} - \mathbf{1}_{\{Y_i \leq t < 0\}}) dt \\ &= \sum_{i=1}^n \int_{-\infty}^{\infty} \mathbb{E}[f'(S^{(i)} + t) | \xi_i = 1] K_i(t) dt. \end{aligned}$$

Since $\sum_{i=1}^n \mathbb{E}X_i^2 = 1$, we have

$$\mathbb{E}f'(S) = \sum_{i=1}^n \int_{-\infty}^{\infty} \mathbb{E}f'(S)K_i(t) dt.$$

Combining two preceding identities we get

$$\begin{aligned} \mathbb{E}(f'(S) - Sf(S)) &= \sum_{i=1}^n \int_{-\infty}^{\infty} \left(Ef'(S) - \mathbb{E}[f'(S^{(i)} + t)|\xi_i = 1] \right) K_i(t) dt \\ &= \sum_{i=1}^n \int_{-\infty}^{\infty} \left(\mathbb{P}(\xi_i = 1) \mathbb{E}[f'(S) - f'(S^{(i)} + t)|\xi_i = 1] \right. \\ &\quad \left. + \mathbb{P}(\xi_i = 0) \left(\mathbb{E}[f'(S)|\xi_i = 0] - \mathbb{E}[f'(S^{(i)} + t)|\xi_i = 1] \right) \right) K_i(t) dt \\ &= \sum_{i=1}^n \int_{-\infty}^{\infty} \left(\frac{k}{n} \mathbf{I} + \left(1 - \frac{k}{n} \right) \mathbf{II} \right) K_i(t) dt. \end{aligned} \tag{11}$$

Next, estimate $|\mathbf{I}|$ and $|\mathbf{II}|$:

$$\begin{aligned} |\mathbf{I}| &\leq \mathbb{E} \left[\left| f'(S^{(i)} + t + (X_i - t)) - f'(S^{(i)} + t) \right| \middle| \xi_i = 1 \right] \\ &\leq \mathbb{E} [\|f''\|_{\infty} |X_i - t| \middle| \xi_i = 1] \leq \|f''\|_{\infty} (\mathbb{E}|Y_i| + |t|), \end{aligned}$$

and

$$\begin{aligned} |\mathbf{II}| &= \left| \mathbb{E}[f'(S^{(i)})|\xi_i = 0] - \mathbb{E}[f'(S^{(i)} + t)|\xi_i = 1] \right| \\ &\leq \left| \mathbb{E}[f'(S^{(i)})|\xi_i = 0] - \mathbb{E}[f'(S^{(i)})|\xi_i = 1] \right| + \|f''\|_{\infty} |t| \\ &= |\mathbf{II}_1 - \mathbf{II}_2| + \|f''\|_{\infty} |t|. \end{aligned}$$

Let us define a random variable J which is independent of (Y_1, \dots, Y_n) and given the vector (ξ_1, \dots, ξ_n) , J is uniformly distributed on $\{j: \xi_j = 1\}$. Note that $\mathcal{L}(S^{(i)} - X_J | \xi_i = 0) = \mathcal{L}(S^{(i)} | \xi_i = 1)$ (both refer to the distribution of the sum of a random sample of $k-1$ random variables out of $Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_n$), thus $\mathbb{E}[f'(S^{(i)} - X_J) | \xi_i = 0] = \mathbf{II}_2$. Hence

$$\begin{aligned} |\mathbf{II}_1 - \mathbf{II}_2| &\leq \mathbb{E} \left[\left| f'((S^{(i)} - X_J) + X_J) - f'(S^{(i)} - X_J) \right| \middle| \xi_i = 0 \right] \\ &\leq \|f''\|_{\infty} \mathbb{E}[|X_J| | \xi_i = 0]. \end{aligned}$$

Since $\mathcal{L}(J | \xi_i = 0)$ is uniform on $\{1, \dots, n\} \setminus \{i\}$ and $\xi_J = 1$ a.s., $\mathbb{E}[|X_J| | \xi_i = 0] = \frac{1}{n-1} \sum_{j \neq i} \mathbb{E}|Y_j|$ thus

$$|\mathbf{II}| \leq \|f''\|_{\infty} \left(|t| + \frac{1}{n-1} \sum_{j \neq i} \mathbb{E}|Y_j| \right).$$

Plugging the bound on I and II into (11) and using (10) and the identity $\mathbb{E}|X_i|^p = \frac{k}{n}\mathbb{E}|Y_i|^p$ (for $p = 1, 2, 3$), we obtain

$$\begin{aligned}
& \mathbb{E}(f'(S) - Sf(S)) \\
& \leq \|f''\|_\infty \sum_{i=1}^n \int_{-\infty}^{\infty} \left(|t| + \frac{k}{n}\mathbb{E}|Y_i| + \left(1 - \frac{k}{n}\right) \frac{\sum_{j \neq i} \mathbb{E}|Y_j|}{n-1} \right) K_i(t) dt \\
& = \|f''\|_\infty \sum_{i=1}^n \left(\frac{1}{2}\mathbb{E}|X_i|^3 + \left(\frac{k}{n}\right)^2 \mathbb{E}|Y_i|\mathbb{E}Y_i^2 + \left(1 - \frac{k}{n}\right) \frac{\sum_{j \neq i} \mathbb{E}|Y_j|}{n-1} \left(\frac{k}{n}\mathbb{E}Y_i^2\right) \right) \\
& \leq \|f''\|_\infty \left(\frac{1}{2} \sum_{i=1}^n \mathbb{E}|X_i|^3 + \left(\frac{k}{n}\right)^2 \sum_{i=1}^n \mathbb{E}|Y_i|^3 \right. \\
& \quad \left. + \frac{k}{n} \left(1 - \frac{k}{n}\right) \frac{1}{n-1} \sum_{i=1}^n \sum_{j \neq i} (\mathbb{E}|Y_j|^3)^{1/3} (\mathbb{E}|Y_i|^3)^{2/3} \right),
\end{aligned}$$

where the last inequality follows from the Hölder inequality. Now, by the standard rearrangement inequality we note that for any integer s ,

$$\sum_{i=1}^n (\mathbb{E}|Y_{i+s}|^3)^{1/3} (\mathbb{E}|Y_i|^3)^{2/3} \leq \sum_{i=1}^n (\mathbb{E}|Y_i|^3)^{1/3} (\mathbb{E}|Y_i|^3)^{2/3} = \sum_{i=1}^n \mathbb{E}|Y_i|^3$$

(the index $i + s$ is taken modulo n and we set $Y_0 = Y_n$). Hence,

$$\sum_{i=1}^n \sum_{j \neq i} (\mathbb{E}|Y_j|^3)^{1/3} (\mathbb{E}|Y_i|^3)^{2/3} = \sum_{s=1}^{n-1} \sum_{i=1}^n (\mathbb{E}|Y_{i+s}|^3)^{1/3} (\mathbb{E}|Y_i|^3)^{2/3} \leq (n-1) \sum_{i=1}^n \mathbb{E}|Y_i|^3.$$

Finally,

$$\mathbb{E}(f'(S) - Sf(S)) \leq \frac{3}{2} \|f''\|_\infty \sum_i \mathbb{E}|X_i|^3.$$

Together with the bound $\|f''\|_\infty \leq 2\|h'\|_\infty$ it yields the estimate for $d_W(S, G)$. To bound $|\mathbb{E}|S| - \sqrt{2/\pi}|$, we take an explicit solution to the equation (8) for $h(x) = |x|$:

$$f(x) = \begin{cases} 1 - 2e^{x^2/2}\Phi(x) & \text{for } x \leq 0, \\ 2e^{x^2/2}(1 - \Phi(x)) - 1 & \text{for } x > 0, \end{cases}$$

where $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt$ is the normal distribution function.

We shall prove that $\|f''\|_\infty = 1$. Note that $f(x)$ is an odd function, thus we compute $f''(x)$ for $x > 0$ only:

$$f''(x) = 2e^{x^2/2}(1 - \Phi(x))(1 + x^2) - x\sqrt{2/\pi}.$$

Since $\lim_{x \rightarrow 0^+} f''(x) = 1$, it suffices to show $f''(x) > 0$ and $f'''(x) < 0$ for all $x > 0$. To this end, we use the estimates for the Gaussian tail proved in [20]: for all $x > -1$,

$$\sqrt{2/\pi} \frac{1}{x + \sqrt{x^2 + 4}} \leq e^{x^2/2} (1 - \Phi(x)) \leq \sqrt{2/\pi} \frac{2}{3x + \sqrt{x^2 + 8}}. \quad (12)$$

We use the lower bound from (12) to prove $f'' > 0$:

$$\sqrt{\pi/2} f''(x) \geq \frac{2(1+x^2)}{x + \sqrt{x^2 + 4}} - x = \frac{x^2 + 2 - x\sqrt{x^2 + 4}}{x + \sqrt{x^2 + 4}}$$

and note that $(x^2 + 2)^2 = x^4 + 4x^2 + 4 > (x\sqrt{x^2 + 4})^2 = x^4 + 4x^2$.

To prove $f'''(x) = 2x(3+x^2)e^{x^2/2}(1 - \Phi(x)) - (x^2 + 2)\sqrt{2/\pi} < 0$ we use the upper bound from (12):

$$2x(3+x^2)e^{x^2/2}(1 - \Phi(x)) \leq \sqrt{2/\pi} \frac{4x(3+x^2)}{3x + \sqrt{x^2 + 8}}.$$

Now it suffices to prove $4x(3+x^2) < (x^2 + 2)(3x + \sqrt{x^2 + 8})$, or equivalently

$$x^3 + 6x < (x^2 + 2)\sqrt{x^2 + 8},$$

which is obvious by calculating $\text{LHS}^2 - \text{RHS}^2 = -32 < 0$. □

Specializing Theorem 13 to the random variables $Y_j = (n/k)^{1/2} \varepsilon_{i,j} v_j$ we arrive with

Proposition 14. *Assume $v \in \ell_2^n$, $\|v\|_2 = 1$, $\|v\|_\infty \leq \alpha$ and let $W = Pv$. Then for all $i = 1, \dots, d$,*

$$|\mathbb{E}|W_i| - \sqrt{2/\pi}| \leq \frac{3}{2} \alpha \sqrt{n/k}.$$

Proof. By (7),

$$|\mathbb{E}|W_i| - \sqrt{2/\pi}| \leq \frac{3}{2} \frac{k}{n} \sum_{j=1}^n \left(\frac{n}{k}\right)^{3/2} |v_j|^3 \leq \frac{3}{2} \alpha \sqrt{n/k}.$$

□

4.4 Concentration of $\|Pv\|_q$ and the proof of the main result

Let $v \in \ell_2^n$ be a unit vector and $W = Pv$. In the two next subsections we shall provide the deviation bounds for $|\|W\|_2^2 - 1|$ and $|\|W\|_1 - \sqrt{2/\pi}|$ as well as combine all the auxiliary results to prove Theorems 2 and 3.

4.4.1 The ℓ_2 case ($q = 2$)

Proposition 15. Assume $v \in \ell_2^n$, $\|v\|_2 = 1$, $\|v\|_\infty \leq \alpha$ and let $W = Pv$. If (6) holds, then for any $t \geq 0$,

$$\mathbb{P}(W_1^2 + \dots + W_d^2 \geq d + t) \leq \exp\left(-\frac{t^2}{6.2d + 12t}\right) + \exp\left(-\frac{3k}{4n\alpha^2} + \log(2d)\right)$$

and

$$\mathbb{P}(W_1^2 + \dots + W_d^2 \leq d - t) \leq \exp\left(-\frac{t^2}{6d}\right).$$

Proof. Denote $Z = W_1^2 + \dots + W_d^2$. Recall that W_1, \dots, W_d are independent, $\mathbb{E}W_i^2 = 1$ thus $\mathbb{E}Z = d$.

First we estimate the upper tail. Let $s_0 = \frac{3}{2} \frac{(k/n)^{1/2}}{\alpha}$. Proposition 10 gives

$$\mathbb{P}(|W_i| \geq s) \leq 2 \exp(-s^2/3) \quad \text{for } 0 \leq s \leq s_0. \quad (13)$$

Set $X_i = W_i^2 \mathbf{1}_{\{|W_i| \leq s_0\}}$ and $\tilde{Z} = \sum_{i=1}^d X_i$. Clearly $\mathbb{E}\tilde{Z} \leq \mathbb{E}Z = d$, hence the union bound and (13) imply

$$\begin{aligned} \mathbb{P}(Z - d \geq t) &\leq \mathbb{P}(\tilde{Z} - d \geq t) + \mathbb{P}(\exists_i |W_i| > s_0) \\ &\leq \mathbb{P}(\tilde{Z} - \mathbb{E}\tilde{Z} \geq t) + 2d \exp(-s_0^2/3). \end{aligned} \quad (14)$$

Next, we estimate $\mathbb{P}(\tilde{Z} - \mathbb{E}\tilde{Z} \geq t)$ using the classical Bernstein inequality for the sum of the variables $X_i - \mathbb{E}X_i$. To this end, we need to verify the condition (4). Note that

$$\mathbb{E}|X_i - \mathbb{E}X_i|^2 = \text{Var}(X_i) \leq \mathbb{E}X_i^2 \leq \mathbb{E}W_i^4 \leq 3.1 \quad (15)$$

where the last inequality follows from Lemma 11. To bound higher moments of $|X_i - \mathbb{E}X_i|$ we use Lemma 12 and the fact that $\mathbb{E}X_i \leq \mathbb{E}W_i^2 = 1$ which imply

$$\mathbb{E}|X_i - \mathbb{E}X_i|^p \leq \mathbb{E}X_i^p + 1 \quad \text{for any } p > 0. \quad (16)$$

By (13), $\mathbb{P}(X_i \geq t) \leq 2 \exp(-t/3)$ for all $t > 0$, hence for any $p > 0$,

$$\mathbb{E}X_i^p = \int_0^\infty pt^{p-1} \mathbb{P}(X_i > t) dt \leq 2 \int_0^\infty pt^{p-1} \exp(-t/3) dt = 2 \cdot 3^p \Gamma(p+1). \quad (17)$$

However, for $p = 3, 4, 5$ we use Lemma 11 to get

$$\mathbb{E}X_i^3 \leq 17, \quad \mathbb{E}X_i^4 \leq 127, \quad \mathbb{E}X_i^5 \leq 1283. \quad (18)$$

Using (15) for $p = 2$, combining (16) with (18) for $p = 3, 4, 5$, and combining (16) with (17) for $p \geq 6$, it is a matter of elementary calculations to verify that

$$\mathbb{E}|X_i - \mathbb{E}X_i|^p \leq \frac{p!}{2} \sigma^2 M^{p-2} \quad \text{for any integer } p \geq 2$$

with $\sigma_i^2 = 3.1$ and $M = 6$. Now the classical Bernstein inequality from Theorem 6 implies

$$\mathbb{P}(\tilde{Z} - \mathbb{E}\tilde{Z} \geq t) \leq \exp\left(-\frac{t^2}{6.2d + 12t}\right)$$

which combined with (14) yields the estimate for the upper tail.

For the lower tail we use Lemma 16 (see below):

$$\begin{aligned} \mathbb{P}(Z - d \leq -t) &= \mathbb{P}\left(\sum_{i=1}^d (1 - W_i^2) \geq t\right) \\ &\leq \exp\left(-\frac{t^2}{2(e^{(2.1)^{-1}} - 1)(2.1)^2 d}\right) \leq \exp\left(-\frac{t^2}{6d}\right), \end{aligned}$$

since $\mathbb{E}(1 - W_i^2)^2 = \mathbb{E}W_i^4 - 2\mathbb{E}W_i^2 + 1 = \mathbb{E}W_i^4 - 1 \leq 2.1$ by Lemma 11. \square

Lemma 16. *Let Y_1, \dots, Y_n are independent random variables, $\mathbb{E}Y_i = 0$, $\mathbb{E}Y_i^2 \leq \sigma^2$ and $Y_i \leq 1$ a.s. Then for any $t > 0$,*

$$\mathbb{P}(Y_1 + \dots + Y_n \geq t) \leq \exp\left(-\frac{t^2}{2(e^{\sigma^{-2}} - 1)\sigma^4 n}\right).$$

Proof. Clearly, we may assume $t \leq n$. Let $a > 0$ to be specified later. Using the inequality

$$e^x \leq 1 + x + \frac{e^a - 1}{a}x^2/2,$$

which is valid for all $x \leq a$, we bound the Laplace transform of Y_i : for any $\lambda \leq a$,

$$\begin{aligned} \mathbb{E}\exp(\lambda Y_i) &\leq \mathbb{E}\left(1 + \lambda Y_i + \frac{e^a - 1}{a}\lambda^2 Y_i^2/2\right) \leq 1 + \frac{e^a - 1}{a}\lambda^2 \sigma^2/2 \\ &\leq \exp\left(\frac{e^a - 1}{a}\lambda^2 \sigma^2/2\right). \end{aligned}$$

Therefore, for any $\lambda \leq a$,

$$\mathbb{P}(Y_1 + \dots + Y_n \geq t) \leq \exp\left(-\lambda t + \frac{e^a - 1}{a}\lambda^2 n \sigma^2/2\right). \quad (19)$$

Taking $\lambda = \frac{t}{\frac{e^a - 1}{a} n \sigma^2}$ and $a = \sigma^{-2}$ we clearly have $\lambda \leq \frac{a}{e^a - 1} \sigma^{-2} = \frac{a^2}{e^a - 1} \leq a$ thus (19) finishes the proof. \square

Proof of Theorem 2. Fix a unit vector $u \in \ell_2^n$. Let $V = HDu$ and set $\alpha := \frac{\sqrt{2e \log(2n/\delta)}}{\sqrt{n}}$. By Proposition 5,

$$\mathbb{P}(\|V\|_\infty \leq \alpha) \leq 1 - \delta. \quad (20)$$

Also, recall $\|V\|_2 = 1$ a.s.

Now, assume $v \in \ell_2^n$ is a fixed unit vector satisfying $\|v\|_\infty \leq \alpha$. We shall show that

$$\mathbb{P}\left(\frac{1}{\sqrt{d}}\|Pv\|_2 \geq 1 + \varepsilon\right) \leq 2\delta/3, \quad \text{and} \quad \mathbb{P}\left(\frac{1}{\sqrt{d}}\|Pv\|_2 \leq \frac{1}{1 + \varepsilon}\right) \leq \delta/3. \quad (21)$$

Since $k \geq 20e \log(2n/\delta)$, (6) is satisfied. For the first inequality in (21), Proposition 15 implies

$$\begin{aligned} \mathbb{P}\left(\frac{1}{\sqrt{d}}\|Pv\|_2 \geq 1 + \varepsilon\right) &\leq \mathbb{P}\left(\|Pv\|_2^2 \geq d + 2d\varepsilon\right) \\ &\leq \exp\left(-\frac{4d\varepsilon^2}{6.2 + 24\varepsilon}\right) + \exp\left(-\frac{3k}{8e \log(2n/\delta)} + \log(2d)\right). \end{aligned}$$

By an elementary calculation one can check that for d and k satisfying the assumptions of the theorem, both exp terms do not exceed $\delta/3$. For the second inequality in (21), Proposition 15 gives

$$\mathbb{P}\left(\frac{1}{\sqrt{d}}\|Pv\|_2 \leq \frac{1}{1 + \varepsilon}\right) \leq \mathbb{P}\left(\|Pv\|_2^2 \leq d\left(1 - \frac{2\varepsilon}{1 + 2\varepsilon}\right)\right) \leq \exp\left(-\frac{2d\varepsilon^2}{3(1 + 2\varepsilon)^2}\right),$$

where the last exp term is $\leq \delta/3$ for d satisfying the hypothesis of the theorem. Thus we proved (21) which implies

$$\mathbb{P}\left(\frac{1}{1 + \varepsilon} \leq \frac{1}{\sqrt{d}}\|Pv\|_2 \leq 1 + \varepsilon\right) \geq 1 - \delta. \quad (22)$$

Finally, since the matrix P and the vector $V = HDu$ are independent, conditioning on V and combining (20) with (22) complete the proof. \square

4.4.2 The ℓ_1 case ($q = 1$)

Proposition 17. Assume $v \in \ell_2^n$, $\|v\|_2 = 1$, $\|v\|_\infty \leq \alpha$ and let $W = Pv$. If (6) holds, then for any $t \geq 0$,

$$\mathbb{P}\left(\left||W_1| + \dots + |W_d| - (\mathbb{E}|W_1| + \dots + \mathbb{E}|W_d|)\right| \geq t\right) \leq 2 \exp\left(-\frac{t^2}{2d + 8t/3}\right).$$

Proof. Denote $Z = |W_1| + \dots + |W_d|$. To estimate $\mathbb{P}(|Z - \mathbb{E}Z| \geq t)$ we use Bernstein inequality (Theorem 6). To this end, we bound the moments of $|W_i|$. By Proposition 10 and integration by parts, for any integer $p \geq 1$,

$$\begin{aligned} \mathbb{E}|W_i|^p &= p \int_0^\infty t^{p-1} \mathbb{P}(|W_i| > t) dt \\ &\leq 2p \int_0^\infty t^{p-1} e^{-t^2/4} dt + 2p \int_0^\infty t^{p-1} \exp\left(-\frac{3t}{4(n/k)^{1/2}\alpha}\right) ds \\ &= p2^p \Gamma(p/2) + 2p! \left(\frac{4}{3}(n/k)^{1/2}\alpha\right)^p. \end{aligned}$$

Plugging the condition (6) we obtain

$$\mathbb{E}|W_i|^p \leq p\Gamma(p/2)2^p + 2p! \left(\frac{4}{3\sqrt{10}} \right)^p.$$

However, for $p \leq 6$ we can provide better bounds. For $p = 4, 6$ we can use Lemma 11 and for $p = 3, 5$ we use the Hölder inequality $\mathbb{E}|W_i|^p \leq \sqrt{\mathbb{E}|W_i|^{p-1}\mathbb{E}|W_i|^{p+1}}$.

Next, Lemma 12 and Proposition 14 combined with (6) imply

$$\begin{aligned} \mathbb{E}||W_i| - \mathbb{E}|W_i||^p &\leq \mathbb{E}|W_i|^p + (\mathbb{E}|W_i|)^p \\ &\leq \mathbb{E}|W_i|^p + \left(\sqrt{2/\pi} + \frac{3}{2\sqrt{10}} \right)^p \leq \mathbb{E}|W_i|^p + (4/3)^p. \end{aligned}$$

Also note that

$$\begin{aligned} \mathbb{E}||W_i| - \mathbb{E}|W_i||^2 &= \mathbb{E}|W_i|^2 - (\mathbb{E}|W_i|)^2 \\ &\leq 1 - \left(\sqrt{2/\pi} - \frac{3}{2\sqrt{10}} \right)^2 \leq \frac{9}{10}. \end{aligned}$$

All these yield the bound

$$\mathbb{E}||W_i| - \mathbb{E}|W_i||^p \leq \frac{p!}{2} \sigma_i^2 M^{p-2}$$

for any integer $p \geq 2$ with $\sigma_i^2 = 1$ and $M = 4/3$, as illustrated by the following table:

p	$\mathbb{E} W_i - \mathbb{E} W_i ^p$	$\frac{p!}{2} \sigma_i^2 M^{p-2}$
2	≤ 0.9	1
3	≤ 3.83	4
4	≤ 5.73	≈ 21
5	≤ 10.6	≈ 142
6	≤ 21.24	≈ 1138
≥ 7	$\leq p\Gamma(p/2)2^p + 2p! \left(\frac{4}{3\sqrt{10}} \right)^p + (4/3)^p$	$\frac{p!}{2} (4/3)^{p-2}$

Now, Bernstein's inequality (Theorem 6) completes the proof. \square

Proof of Theorem 3. As in the proof of Theorem 2, it is enough to show

$$\mathbb{P} \left(\left| \frac{1}{d} \|Pv\|_1 - \sqrt{2/\pi} \right| \geq \varepsilon \sqrt{2/\pi} \right) \leq \delta \quad (23)$$

for any unit $v \in \ell_2^n$ satisfying $\|v\|_\infty \leq \alpha := \sqrt{2e \log(2n/\delta)}/\sqrt{n}$.

Fix $\kappa \in (0, 1)$. The condition (6) is satisfied since $k \geq 20e \log(2n/\delta)$. Proposition 17 used for $t = d\kappa\varepsilon\sqrt{2/\pi}$ implies

$$\mathbb{P} \left(\left| \frac{1}{d} \|Pv\|_1 - \mathbb{E} \|Pv\|_1 \right| \geq \kappa\varepsilon\sqrt{2/\pi} \right) \leq 2 \exp \left(- \frac{\kappa^2(2/\pi)d\varepsilon^2}{2 + \sqrt{2/\pi}\frac{8}{3}\kappa\varepsilon} \right) \leq \delta,$$

where the last inequality follows from the assumption on d , while Proposition 14 yields

$$\left| \frac{1}{d} \mathbb{E} \|P_v\|_1 - \sqrt{2/\pi} \right| \leq \frac{3\sqrt{2e \log(2n/\delta)}}{2\sqrt{k}} \leq (1 - \kappa)\varepsilon\sqrt{2/\pi},$$

where the last inequality follows from the assumption on k . \square

Acknowledgments: A preliminary version of this work was presented at the conference “Perspectives in High Dimensions”, held at Case Western Reserve University in August 2010. The author would like to thank Prof. Louis H. Y. Chen for a discussion on normal approximation problem encountered in this work (Theorem 13).

References

- [1] D. Achlioptas, Database-friendly random projections: Johnson-Lindenstrauss with binary coins, *J Comput System Sci*, 66 (2003), 671–687.
- [2] N. Ailon and B. Chazelle, Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform, *Proc 38th ACM Symp Theory of Computing*, 2006, pp. 557–563.
- [3] N. Ailon and E. Liberty, Fast dimension reduction using Rademacher series on dual BCH codes. *Discrete Comput Geom* 42 (2009), no. 4, 615–630.
- [4] N. Alon, L. Babai, and A. Itai, A fast and simple randomized parallel algorithm for the maximal independent set problem, *J Algorithms* 7 (1986), no. 4, 567–583.
- [5] E. Bolthausen, An estimate of the remainder in a combinatorial central limit theorem, *Z Wahrsch Verw Gebiete* 66 (1984), no. 3, 379–386.
- [6] A. Bonami, Étude des coefficients de Fourier des fonctions de $L^p(G)$, *Ann Inst Fourier (Grenoble)* 20 (1970), 335–402.
- [7] L. H. Y. Chen and Q. M. Shao, Stein’s method for normal approximation, in: *An introduction to Stein’s method*, *Lect Notes Ser Inst Math Sci Natl Univ Singap*, 4, Singapore Univ Press, Singapore, 2005, pp. 1–59.
- [8] L. H. Y. Chen and X. Fang, On the error bound in a combinatorial central limit theorem, preprint, [arXiv:1111.3159v1](https://arxiv.org/abs/1111.3159v1)
- [9] S. Dasgupta and A. Gupta, An elementary proof of a theorem of Johnson and Lindenstrauss, *Random Structures Algorithms* 22 (2003), 60–65.
- [10] A. Hinrichs and J. Vybíral, Johnson-Lindenstrauss lemma for circulant matrices, *Random Structures Algorithms* 39 (2011), 391–398.
- [11] S. T. Ho and L. H. Y. Chen, An L_p bound for the remainder in a combinatorial central limit theorem, *Ann Probab* 6 (1978), no. 2, 231–249.

- [12] W. Hoeffding, Probability inequalities for sums of bounded random variables, *J Amer Statist Assoc* 58 (1963), 13–30.
- [13] P. Indyk and R. Motwani, Approximate nearest neighbors: Towards removing the curse of dimensionality, *Proc 30th Annu ACM Symp Theory of Computing*, Dallas, TX, 1998, pp. 604–613.
- [14] K. Joag-Dev and F. Proschan, Negative association of random variables, with applications, *Ann Statist* 11 (1983), 286–295.
- [15] W. B. Johnson and J. Lindenstrauss, Extensions of Lipschitz mappings into a Hilbert space, *Contemp Math* 26 (1984), 189–206.
- [16] D. M. Kane, R. Meka, and J. Nelson, Almost optimal explicit Johnson-Lindenstrauss transformations, *Proc 15th Int Workshop on Randomization and Computation* 2011.
- [17] J. Matoušek, On variants of the Johnson-Lindenstrauss lemma, *Random Structures Algorithms* 33 (2008), 142–156.
- [18] F. J. MacWilliams and N. J. A. Sloane, *The theory of error-correcting codes*, North-Holland Mathematical Library, Vol. 16. North-Holland Publishing Co., 1977.
- [19] Q. M. Shao, A comparison theorem on moment inequalities between negatively associated and independent random variables, *J Theoret Probab* 13 (2000), 343–356.
- [20] S. J. Szarek and E. Werner, A nonsymmetric correlation inequality for Gaussian measure, *J Multivariate Anal* 68 (1999), no. 2, 193–211.
- [21] J. Vybíral, A variant of the Johnson-Lindenstrauss lemma for circulant matrices, *J Funct Anal* 260 (2011), 1096–1105.